ANALYSE DES DONNÉES

L'ADD est interreliée avec 3 autres disciplines (la statistique, l'informatique, l'étude)

Les types d'étude

Expérimentales :

Réalité observée telle qu'elle se présente (études d'observation)

Recherche et développement

Test de produit, Test de marché.

Non expérimentale :

Manipulation de l'exposition au facteur étudié pour ensuite observer l'effet

Qualitatives : Entretiens – Observation – Cas – chainage Cognitif

Quantitative : Questionnaires...

Entretiens :

Monadique : Une seul sens

Dyadique : Approche transactionnelle – interactive (relation clt produit)

Non directifs :

Semi directifs : Problématique non claire

Taille de l'échantillon : 30 au minimum :

Ne doit pas être représentatif au sens statistique (cov=0) mais seulement au sens théorique

Saturation théorique : Chaque fois qu'un entretien supplémentaire n'apporte pas d'informations nouvelles, toute fois l'échantillon doit être hétérogène sur le plan sociodémographique.

ADD : Critères de ressemblance > Triage > Classification des données suivant une approche logique.

On applique l'analyse du contenu une fois les entretiens réalisés, les verbatives doivent être générés (la retranscription mot à mot des propos d'un entretien)

L'analyse du contenu se distingue en deux, lexicale et syntaxique [l'analyse porte sur le mot]

Et l'analyse thématique [associations de mots]

Magasinage : quête des nouvelles

Browsing Butinage : aller en magasin sans avoir l'intention d'achat.

L'analyse peut être faite de façon horizontale ou verticale, la première porte sur le verbative pris un à un, alors que la 2d établie un regroupement entre tous les verbatives.

Les entretiens peuvent être individuels ou en groupe.

Chainage cognitif : Analyse de la personnalité.

Méthodes projectives : le premier mot qui vous passe par la tête.

Etude de cas : Mono cas ou multi cas.

Pour réussir la conduite des entretiens il faut inspirer la confiance avec les répondants, cette confiance constitue alors la pierre angulaire pour que le répondant ne fait pas recours aux mécanismes de défense, ces derniers sont en nombre de 4 :

La non réponse ou le refuge dans les mensonges

La rationalité : approche normative

L'imputation sur autrui

Le refoulement : fuir la réponse.

Les études qualitatives : Variables subjectives

Mesurer (opérationnaliser) les Variables abstraites : Latentes, non observables

Satisfaction = Cognitive + Affective

Engagement = Affectif + Calculé

Fidélité = Comportementale + d'attitude

Echelles :

Nominales : non métrique mesurent l'appartenance à une classe.

Ex: Le montant de vos achats a-t 'il augmenté ces trois derniers mois? - oui – non

Ordinales: échelles de classement qui établissent une relation d'ordre entre des objets et cela par rapport à un critère de classement prédéfini.

Ex: classement de produits par ordre de préférence, de qualité perçue

Variable d'intervalle: il s'agit d'une échelle métrique dont les unités de mesure sont constantes et pour lesquelles les distances entre niveaux sont connus.

L'origine d'une échelle d'intervalle reste toutefois arbitraire.

Variable de rapport: dans ce type d'échelle il existe une unité de mesure prédéfinie (kg,cm, F...) et un zéro naturel qui correspond à l'absence du phénomène étudié.

Satisfaction factorielle > Satisfaction relationnelle > Confiance > Fidélité > engagement

La validité

La validité du contenu : les questions doivent traduire et répondre à la problématique et aux objectifs de l'étude. (deux experts professionnels et deux linguistes)

La validité du trait : Degré auquel on peut affirmer que le construit opérationnalisé permet de mesurer le concept qu'il est censé représenter

Validité convergente : concerne la capacité d'un test à pointer les sujets dans les catégories réalisées.

Validité discriminante : La validité discriminante vérifie l'envers de la validité convergente: il s'agit de savoir si seul le construit mesuré est mesuré par le test. En d'autres termes, il s'agit de savoir si le test fait bien la différence entre le construit mesuré et n'importe quel autre.

Etudes Marketing

Echantillon De 5 à 6 fois le nombre d'items

Aléatoire simple : chaque membre d'une population a une chance égale d'être inclus à l'intérieur de l'échantillon

Aléatoire stratifié : on divise la population en groupes homogènes (appelés strates), qui sont mutuellement exclusifs, puis on sélectionne à partir de chaque strate des échantillons indépendants.

1. Segmenter la population en strates
2. Un échantillon pour chaque strate

Aléatoire systématique : signifie qu'il existe un écart, ou un intervalle, entre chaque unité sélectionnée qui est incluse dans l'échantillon.

Échantillonnage en grappes

La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou en grappes comme son nom l'indique. Suivant cette technique, on sélectionne au hasard un certain nombre de grappes pour représenter la population totale, puis on englobe dans l'échantillon toutes les unités incluses à l'intérieur des grappes sélectionnées.

Redresser un échantillon

Dans le domaine des études marketing, le redressement d'échantillons a pour objectif d'améliorer la représentativité de l'échantillon interrogé, sur un certain nombre de critères de qualification. Le principe sous-jacent est que seul un échantillon ayant la même structure que la population-mère sur les critères que l'on connaît de cette population, permet de généraliser les réponses obtenues sur les autres critères, à l'ensemble de cette population. Le redressement cherche donc à appliquer des pondérations aux individus pour augmenter le poids de ceux appartenant à des groupes sous-représentés dans l'échantillon interrogé par rapport à la population-mère, et à réduire parallèlement le poids de ceux qui sont surreprésentés.

Prévision : données du passé

Prédiction : données actuelles

Tri à plat : une seule variable

Tri croisé : plusieurs variables

Types d'études :

Echelles	Type d'étude
Nominales	Descriptive
Ordinales	Explicative
Intervalle	Prédictive
Rapport	causale

Etude descriptive : Décrire une population à travers des variables choisis /échelle nominales

Etudes explicatives : expliquer un phénomène

Variables dépendantes 'à expliquer : endogène'

Une ou plusieurs variables indépendantes : explicatives ou endogènes

Etude prédictive : hypothèse > vérification de sa validité

Analyse discriminante (échelles d'intervalle, ratio) > Analyse typologique

Etude causale : Variable médiatrice, modulatrice

Type d'analyse :

Univariée : une variable

Bivariée : deux variables (une à expliquer – explicative)

Multivariée : (variable à expliquer – plusieurs variables explicatives)

Variables métrique : Tests paramétriques

Variables non métrique : tests non paramétriques

Tests paramétriques

Supposent que les variables sont d'intervalle ou de rapport et qu'elles sont distribués selon une loi normale

Lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres. L'hypothèse de normalité sous-jacente des données est le plus souvent utilisée, la moyenne et la variance suffisent pour caractériser complètement la distribution. Concernant les tests d'homogénéité par

exemple, pour éprouver l'égalité des distributions, il suffira de comparer les moyennes et/ou les variances.

Les tests non paramétriques

S'appliquent à des variables nominales ou ordinales, ils n'exigent pas que les données soient distribuées d'une façon particulière.

Lorsque l'on dispose d'un seul échantillon les tests les plus utilisés sont le test de Kolmogorov-Simirnov et le text de chi-deux.

Ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests distribution free. L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire.

Lorsque les données sont quantitatives, les tests non paramétriques transforment les valeurs en rangs. L'appellation tests de rangs est souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

La distinction paramétrique – non paramétrique est essentielle. Elle est systématiquement mise en avant dans la littérature. Les tests non paramétriques, en ne faisant aucune hypothèse sur les distributions des données, élargissent le champ d'application des procédures statistiques. En contrepartie, ils sont moins puissants lorsque ces hypothèses sont compatibles avec les données.

TESTS paramétriques :

T Student et le test F

Permettent de comparer la moyenne des réponses de l'échantillon à celle estimée dans la population mère ou à une valeur théorique connue.

T calculé = SCE/1 / SCR/n-2

Coefficient de corrélation: R : Mesure de l'association entre deux variables sur une échelle d'intervalle ou de rapport.

TESTS non paramétriques :

Entre variables nominales :

Test de chi-deux : mesurer l'ajustement de la distribution des fréquences d'une variable nominale ou bien l'association entre deux variables nominales extraites d'échantillons indépendants.

1. Une seule variable :

Savoir si les fréquences observées sont différentes de de celles estimées dans la population.

Calcul des écarts pour tester la probabilité qu'ils se produisent sous l'hypothèse nulle Ho qui postule l'égalité des distributions.

X^2= Sigma 1>k de $(Oi- Ti)^2 / Ti$ O : fréquence observée ; T : Fréquence théorique k : nombre de catégories

Si $X^2=0$ les deux variables sont parfaitement indépendantes

Plus X^2 est rand plus il est probable que les deux variables soient dépendants (on n'est plus sur)

Test de Kolmogorov-Smirnov

Test d'ajustement ; comparer une distribution observée à une distribution théorique

= Max (Propo cumulée observée – Prop cumulé théorique)

Analyse multi variée

Explorer la structure d'une base de données, ou identifier les relations entre les variables appartenant à cette base.

Méthodes exploratoires :

L'exploration de la structure d'une base de données.

- Analyse factorielle en composantes principales (échelles d'intervalle)
- Analyse des similarités et des préférences (échelles ordinales)
- Analyse factorielle des correspondances (échelles nominales)
- Analyse des sujets (Analyse typologique) – pour toutes les échelles

Méthodes explicatives

Apres avoir identifier la variable à expliquer et les variables explicatives

Tests de normalité

Ils permettent de confirmer une normalité.
Permettent de vérifier si des données réelles suivent une loi normale ou non. Les tests de normalité sont des cas particuliers des tests d'adéquation (ou tests d'ajustement, tests permettant de comparer des distributions), appliqués à une loi normale.

Test de Shapiro Wilk

Il compare une distribution observée à une loi Gaussienne.
H0 : La répartition observée est compatible avec la normalité.

Test de Kolmogorov-Smirnov

Les distributions doivent être continues. Meilleur que le précédent si les effectifs sont gros (n>2000).
Basé sur la loi forte des grands nombres (fonction de répartition empirique).
H0 : G1=G2

Les Coefficients d'asymétrie et d'aplatissement sont également utiles pour définir une loi normale.

Pour l'aplatissement : le degré de concentration des observations

$$G_2 = \frac{(n+1)\,n}{(n-1)\,(n-2)\,(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 - 3\,\frac{(n-1)^2}{(n-2)(n-3)}$$

Et pour l'asymétrie : la symétrie de la distribution des réponses autour de la valeur centrale

$$G_1 = \frac{n}{(n-1)\,(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3$$

Avec σ est la racine d'un estimateur non biaisé de la variance.

Test d'ajustement et d'interférence :

Ajustement : vérifier l'ajustement de la distribution des fréquences ou de la moyenne à une distribution théorique ou à une moyenne connue

Interférence : variance, régression : la relation particulière qu'entretiennent deux variables ou plus de deux variables

ANALYSE FACTORIELLE EXPLORATOIRE ET CONFIRMATOIRE

L'analyse factorielle cherche à réduire un nombre important d'informations (prenant la forme de valeurs sur des variables) à quelques grandes dimensions. Comme dans toute analyse statistique, on tente donc d'expliquer la plus forte proportion de la variance (de la covariance dans le cas de l'analyse factorielle) par un nombre aussi restreint que possible de variables (appelées ici composantes ou facteurs). On utilise le terme de variables latentes pour parler de ces variables qui existent au plan conceptuel seul et qui ne sont pas mesurées.

ANALYSE FACTORIELLE EXPLORATOIRE - ACP

Identifier un ensemble de dimensions latentes (non observables) à partir d'un seul ensemble plus grand de variables initiales, il s'agit alors de découvrir une structure sous-jacente (nature et nombre de dimensions)

Par exemple : à partir de 255 questions posées dans le test de personnalité 15 dimensions peuvent être identifiées qu'on appelle facteurs.

Purification des données > transformation des données non métriques et D métriques

L'analyse factorielle en composantes simples : la construction d'échelles destinées à mesurer des caractéristiques individuelles de consommateur ou d'entreprises :

Par exemple : 14 items mesurant la confiance envers la marque peuvent se réduire en trois dimensions ; la crédibilité, l'intégrité et la bienveillance.

Analyse de la matrice de corrélations ou de covariances :

Si les variable étudiées sont indépendantes les unes des autres l'analyse factorielle ne sert à rien car elle fournira autant de facteurs que de variables, il est donc dans ce cas impossible de résumer l'information

Matrice de corrélation sou matrice de covariance : l'analyse factorielle permet d'effectuer une classification automatique ou typologique.

Il est préférable d'utiliser la matrice de covariance lorsqu'on envisage une comparaison des structures factorielles entre groupes.

Pour déterminer si les corrélations existantes sont suffisantes pour effectuer une analyse factorielle on utiliser les3 indicateurs : Test de sphéricité de Barlett, KMO, MSA

Choix d'une analyse factorielle exploratoire :

L'analyse doit ensuite choisir entre l'analyse en composantes principales ACP et l'analyse en facteurs communs AFC la différence repose sur la nature des facteurs

ACP : facteurs : indices formés par les variables (indicateurs formulatifs des composantes)

Appropriée lorsque l'on cherche à prédire les scores des sujets des facteurs, calculer les indices, réduire l'ensemble des variables ou encore expliquer la variance

AFC : variables : reflet de facteurs latents (indicateurs réflectifs)

Lorsqu'on cherche à mettre en évidence des dimensions ou des construits latents dont les variables observées ne sont que le reflet et qu'on cherche à éliminer l'erreur ou la variance spécifique contenue dans chaque variable

Validation de l'analyse factorielle :

La généralisation des résultats obtenus sur la population nécessite une réplique sur un groupe tiré aléatoirement de la population

Pour vérifier l'identité des structures entre deux analyses factorielles exploratoires ; on peut effectuer une analyse factorielle confirmatoire ; celle-ci permet la comparaison de la structure factorielle obtenue entre les deux groupes.

Analyse confirmatoire : Analyse + confirmation du modèle

Transformation en variable métrique (base virtuelle)

Purifier

Analyse factorielle exploratoire : faire émerger une théorie et concevoir un modèle théorique

Analyse factorielle confirmatoire : mettre à l'épreuve des hypothèses spécifiques concernant l'influence des variables latentes sur les données recueillies ; elle permet de tester un modèle théorique.

Analyse Marketing :

Etape 1 : Test de normalité des données : Réponses étalées = concentration des données

La loi des grands nombres : loi normale

Choix d'un grand échantillon pour normalisé les données

Bootstart : Étirassions (essais) augmenter la finalisation

Test d'asymétrie < |2|

Test d'aplatissement < |7|

Factorisabilité des données

Analyse des données > Données factorisables

TEST KMO > 0,5

Loading ou qualité de représentaiton des items

2. Calcul de la variance expliquée : min=60%

Rotation diagonale

Rotation anti diagonale

Influence mutuelle des items

Alpha de Cronbach

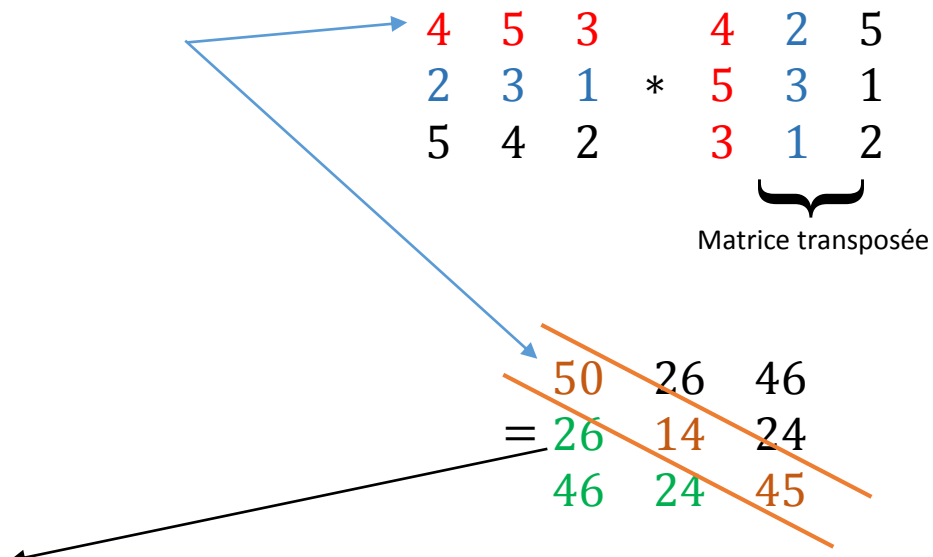
Plus la valeur alpha s'approche de 1, plus l'ensemble d'éléments est homogène.

$$\left(\frac{k}{k-1}\right)\left(1-\frac{\sum Vei}{\sum Vi+2Cov\ ij}\right.$$

Inconvénient : sensible au nombre d'items k

Exemple :

	Item 1	Item2	Item3
Réponse 1	4	5	3
Réponse 2	2	3	1
Réponse 3	5	4	2



50=4*5+5*5+3*3 (1ère ligne matrice initiale * 1ère colonne matrice transposée)

26=2*4+3*5+1*3 (2ème ligne matrice initiale * 1ère colonne matrice transposée)

…. And so on

$$=\left(\frac{3}{3-1}\right)\left(1-\frac{109}{109+2*96}\right.=0{,}96$$

109= 50+14+45 96=26+46+24

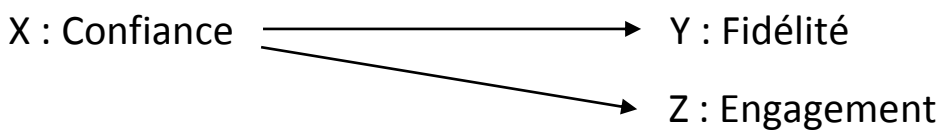
Rho de Joreskog : Validité convergente :

résout le problème de sensibilité au nombre d'items

$$\sum alpha\ i^2 / \sum alpha\ i^2 + \sum Vai$$

Validité prédictive : Résultat prévu du terrain

Test de médiation



1ère condition : Relier x avec z par le billet de la régression linéaire simple, on vérifie la signification des éléments suivants : R² ajusté, coefficient standardisé B, T de student

A préciser R²a dépend de la taille de l'échantillon (Ts> 0,96 pour un seuil de 5%)

2ème condition : On régresse y par rapport à x et on applique la même logique .

3ème condition : On régresse z par rapport à x et y

1. Tester les relations (Test de Student)
2. La nature du médiateur (partiel ou total) (Test de Sobel)

Médiateur total : Si l'influence de x/z disparait en présence de y
Si non on parle de médiateur partiel

3. Calcul de la part de l'effet médiateur par rapport à l'effet total
   Combien de Y pour avoir Z

   Validité du contenu du questionnaire > Collecte des données

Rho de Joreskog : Validité convergente

Validité prédictive : Résultat prévus = Résultats sur le terrain (?)